OnForumS: A Shared Task on On-line Forum Summarisation

Mijail Kabadjov,
Udo Kruschwitz,
Massimo Poesio
University of Essex
Colchester, UK
{malexa,udo,poesio}
@essex.ac.uk

Josef Steinberger
University of West Bohemia
Pilsen, Czech Republic
jstein@kiv.zcu.cz

Emma Barker
University of Sheffield
Sheffield, UK
e.barker@sheffield.ac.uk

Abstract

In this paper we present the Online Forum Summarisation (OnForumS) pilot task at MultiLing'15. OnForumS is a pioneering attempt at encompassing automatic summarisation, argumentation mining and sentiment analysis into one shared task and at bringing crowdsourcing to the evaluation of systems for automatic summarisation and argument structure parsing. Four research groups, each submitting two runs, participated in the task and these complemented with two baseline system runs were evaluated via crowdsourcing. Performance results are presented and briefly discussed. Being the first of its kind, we believe OnForumS'15 was a successful campaign and hope it will establish itself as a valuable exercise in advancing the state-of-the-art in this new emerging area.

1 Introduction

Most major on-line news publishers, such as The Guardian or Le Monde, publish articles on different topics and encourage reader engagement through the provision of an on-line comment facility. A given news article can often give rise to thousands of reader comments – some related to specific points within the article, others that are replies to previous comments. The high volume of such user-supplied comments suggests the need for automated methods to summarise this content, which in turn poses an exciting and novel challenge for the summarisation community.

The problem of producing a digest of such mass of comments touches on at least three areas of research in Natural Language Processing, as are Automatic Summarisation (Erkan and Radev, 2004; Giannakopoulos and Karkaletsis, 2013), Argumentation Mining (Palau and Moens, 2011; Boltuzic and Šnajder, 2014) and Sentiment Analysis (Pang and Lee, 2008; Turney and Littman, 2003).

The Online Forum Summarisation (OnForumS) pilot task at MultiLing'15[1] is a pioneering attempt at encompassing all three areas into one shared task in order to investigate how the mass of comments found on news providers' web sites can be summarised. We posit that a crucial initial step towards that goal is to determine what comments link to, be that either specific points within the text of the article, the global topic of the article, or comments made by other users. This constitutes a linking task. Furthermore, a set of types or labels for a given link may be articulated to capture phenomena such as agreement (e.g., in favour, against) and sentiment (e.g., positive or negative) with respect to the comment target.

The main contribution of this paper is two-fold: firstly, the operationalisation of this labelled linking task into a shared task – to our knowledge the first of its kind – and secondly, casting the evaluation of such task as a crowdsourcing campaign – using crowdsourcing for the evaluation of both summarisation and argument structure has been largely under-explored in previous work.

The remainder of the paper is organised as follows, section 2 describes the shared task and the data set collection and preparation, section 3, discusses results and covers the sampling and evaluation strategy harnessing crowdsourcing, section 4 provides a brief literature survey and finally conclusions are drawn with pointers to future work.

2 On-line Forum Summarisation

The On-line Forum Summarisation (OnForumS) is a particular specification of the linking task

[1] http://multiling.iit.demokritos.gr/pages/view/1516/multiling-2015

mentioned in the previous section, in which systems take as input a news article with associated comments[2] and are expected to link each comment sentence to article sentences (which, for simplification, are assumed to be the appropriate units here) or to preceding comments and then to label each link for argument structure $in\_favour, against, impartial$ and sentiment $positive, negative, neutral$.[3] Data for the task is collected in two languages English and Italian.[4]

Evaluation of systems output is based on the results of a crowdsourcing exercise, which although widely used in other areas such as Machine Translation (Callison-Burch, 2009), Opinion Mining (Snow et al., 2008) and Word Sense Disambiguation (Passonneau and Carpenter, 2013), to a much lesser extent it has been employed for evaluating Summarisation and never before in previous MultiLing campaigns. In our case, contributors are asked to judge whether potential links and associated labels are correct for each test article and its comments. The crowdsourcing HIT is defined as a validation task as opposed to annotation, that is, contributors are only asked to validate links and labels produced by systems and are not asked to link or label data themselves. Additionally, due to the high volume of system links only a subset of all the links produced by systems is evaluated by extracting a stratified sample.

2.1 Defining the task

Linking comment sentences to article sentences is a useful step towards summarising the mass of comments. For instance, comment sentences linked to the same article sentence can be seen as forming a "cluster" of sentences on a specific point/topic. Moreover, having labels capturing argument structure and sentiment enables computing statistics within such topic clusters on how many readers are in favour or against the point raised by the article sentence and what is the general 'feeling' about it. Consider the following example from our corpus:

Table 1: OnForumS Corpus.

Concept	**English**	**Italian**
Number of words	43104	34803
Links validated (via crowdsourcing)	2311	1087
All Links	9635	6193
Unique Links and Labels	6576	4138
Unique Links only	5789	4016
Type d Links	3517	2083
Type c Links	2975	2024
Type b Links	63	20
Type a Links	21	11

(1) **AS:** In September the environment secretary, Owen Paterson, assured us that climate change "is something we can adapt to over time and we are very good as a race at adapting".
↪ **C1:** Human adaptability!!!!!!!!!!!!!! Tell that to ther first dynasty of Egypt (the ones with the pyramids), who died from hunger due to a 30-year drought, the Minoans (volcanic eruption and tsunami), Babylonians (drought), ...
→ **C2:** Patronising and cynical comment by the Government. I daresay we can 'adapt' to a certain extent but there are limits.

In example 1, the first comment ($C1$) links to article sentence AS through 'human adaptability' and it expresses a view against the quote given in AS and then the second comment ($C2$) seconds the viewpoint of $C1$ (it is actually a reply to $C1$).

Such clusters of linked sentences are not summaries in themselves, but can be seen as digests of the mass of comments and key points covered in news articles (to an extent resembling the idea of 'capsule overview' put forward in (Boguraev and Kennedy, 1997)).

2.2 Data

Data for the task was collected in English and Italian. A sample data consisting of one article in English and small set of comments and labelled links result of internal pre-pilots was released early on. The official test data set consisted of ten articles from The Guardian (EN) and six articles from La Repubblica (IT) together with corresponding top fifty comments for each article (see Table 1).

3 Evaluation via Crowdsourcing

Four research groups participated in the OnForumS shared task, each group submitting two runs. In addition, two baseline system runs were included making a total of ten different system runs.

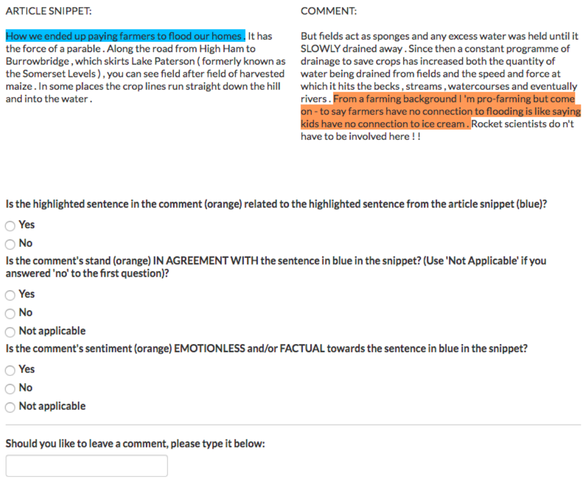
Submissions were evaluated via crowdsourcing[5], which is a commonly used method for eval-

[2] Only the top fifty comments filtered according to number of likes and number of replies are included (articles may contain thousands of comments).

[3] The search space for links is defined by the union of Cartesian products of article sentences with comment sentences and comment sentences with comment sentences: $AS \times CS \cup CS \times CS$).

[4] Sample and test data for the task were released in an XML format pre-tokenised and sentence-split (see `http://multiling.iit.demokritos.gr/pages/view/1531/task-onforums-data-and-information`).

[5] We used CrowdFlower: `http://www.`



ARTICLE SNIPPET:

How we ended up paying farmers to flood our homes . It has the force of a parable . Along the road from High Ham to Burrowbridge , which skirts Lake Paterson (formerly known as the Somerset Levels) , you can see field after field of harvested maize . In some places the crop lines run straight down the hill and into the water .

COMMENT:

But fields act as sponges and any excess water was held until it SLOWLY drained away . Since then a constant programme of drainage to save crops has increased both the quantity of water being drained from fields and the speed and force at which it hits the becks , streams , watercourses and eventually rivers . From a farming background I 'm pro-farming but come on - to say farmers have no connection to flooding is like saying kids have no connection to ice cream . Rocket scientists do n't have to be involved here ! !

Is the highlighted sentence in the comment (orange) related to the highlighted sentence from the article snippet (blue)?

○ Yes
○ No

Is the comment's stand (orange) IN AGREEMENT WITH the sentence in blue in the snippet? (Use 'Not Applicable' if you answered 'no' to the first question)?

○ Yes
○ No
○ Not applicable

Is the comment's sentiment (orange) EMOTIONLESS and/or FACTUAL towards the sentence in blue in the snippet?

○ Yes
○ No
○ Not applicable

Should you like to leave a comment, please type it below:

Figure 1: Validation HIT on CrowdFlower.

uating HLT systems (Snow et al., 2008; Callison-Burch, 2009). The crowdsourcing HIT was designed as a validation task (as opposed to annotation), where each system-proposed link and labels are presented to a human contributor for their validation with both article sentence and comment sentence placed within context (see Fig. 1).

Both the HIT and the instructions for contributors were translated to English and Italian, thus targeting two distinct groups of native speakers.

3.1 OnForumS Evaluation

The approach used for the OnForumS evaluation is IR-inspired and based on the concept of *pooling* used in TREC (Soboroff, 2010), where the assumption is that possible links that were not proposed by any system are deemed irrelevant. Then from those links proposed by systems, four categories are formed as follows (see Table 1 for the cumulative distribution of each):

(a) links proposed in 4 or more system runs
(b) links proposed in 3 system runs
(c) links proposed in 2 system runs
(d) links proposed only once

Due to the volume of links proposed by systems, a stratified sample was extracted for evaluation based on the following strategy: all of the **a** and **b** links[6], one third of the **c** links selected at random and one third of the **d** links also selected at random (see Table 1 for numbers of links validated via crowdsourcing).

Once the crowdsourcing exercise was completed, correct and incorrect links were counted first for the linking task only based on the aggregated judgements provided by Crowd Flower[7] (i.e., number of 'yes' and 'no' answers from contributors). From those links validated as correct, the correct and incorrect argument and sentiment labels were counted (again, number of 'yes' and 'no' answers). Using these counts precision scores were computed and system runs were then ranked based on these precision scores. For the linking task no system surpassed the baseline algorithm based on overlap but scores were substantially higher for English than for Italian (see Tables 2 and 3).

3.2 Creating a Gold Standard to Estimate Recall

There are two ways to create gold standard links and labels from the validated data. One is direct validation which entails taking all 'yes' validations of links as gold links and then all labels for argument and sentiment with 'yes' validations as the gold labels for those links. And the other way is by exclusion, if all possible labels for a given link except for one have a 'no' validation then this makes the remaining label a gold label (e.g., if it is not "against", nor "impartial", then it is "in_favour"). With these criteria in mind we created a small gold standard set from which precision, recall and F1 can be computed (see Table 4).[8]

From Table 4 we can see that recall ranged between 45 70% and precision, 24 25%, for the labels *In_Favour* and *Positive*, and precision, 3 5% and around 5% for labels *Against* and *Negative*, respectively. A visualisation of systems performance in terms of precision/recall scatter plots can be included in camera-ready version (omitted here due to space constraints).

crowdflower.com

[6] The popular links (**a** and **b**) were not that many, hence, we chose to include all.

[7] An aggregated judgement is based on multiple judgements using CrowdFlower's "agg" method which returns a single "top" result – AKA the contributor response with the highest confidence (agreement weighted by contributor trust) for every given data point (for more details see: `https://success.crowdflower.com/hc/en-us/articles/203527635-CML-and-Instructions-CML-Attribute-Aggregation`).

[8] We include P/R/F1 measures only for English as for Italian the number of 'yes' responses was substantially smaller, and hence, the gold set of labels too.

Table 2: System Ranking according to Precision: English.

System-run	Linking	System-run	Argument	System-run	Sentiment
BASE-overlap	93.1	CIST-run2	99.3	CIST-run1	95.1
USFD_UNITN-run2	88.7	CIST-run1	99.1	CIST-run2	93.9
UWB-run1	86.5	UWB-run1	97.5	BASE-overlap	93.8
UWB-run2	86.5	UWB-run2	97.5	BASE-first	93.5
JRC-run1	86.2	BASE-first	92.7	USFD_UNITN-run2	92.6
JRC-run2	83.1	JRC-run2	90.7	JRC-run2	90.3
USFD_UNITN-run1	81.9	USFD_UNITN-run1	89.4	USFD_UNITN-run1	89.8
BASE-first	74.3	JRC-run1	88.9	UWB-run1	88.9
CIST-run2	71.8	BASE-overlap	88.6	UWB-run2	88.9
CIST-run1	70.9	USFD_UNITN-run2	86.2	JRC-run1	87.9

Table 3: System Ranking according to Precision: Italian.

System-run	Linking	System-run	Argument	System-run	Sentiment
BASE-overlap	59.1	CIST-run2	1	CIST-run1	66.7
UWB-run1	25	UWB-run1	1	BASE-overlap	50
USFD_UNITN-run1	20	CIST-run1	77.8	JRC-run1	37.5
JRC-run1	15.2	BASE-first	75	BASE-first	33.3
CIST-run1	8.4	BASE-overlap	69.2	UWB-run1	25
CIST-run2	3.3	JRC-run1	44	CIST-run2	0
BASE-first	1.0	USFD_UNITN-run1	0	USFD_UNITN-run1	0

4 Related Work

Producing a digest of the mass of comments found on news providers' web sites with their associated news article content lies at the intersection of three areas of research in Natural Language Processing, as are Automatic Summarisation, Argumentation Mining and Sentiment Analysis. Whilst the former has been an active area of research for decades (Luhn, 1958; Boguraev and Kennedy, 1997; Erkan and Radev, 2004; Giannakopoulos and Karkaletsis, 2013), the latter two are newer areas that have gained much interest in recent years (Palau and Moens, 2011; Pang and Lee, 2008).

A good literature survey on Automatic Summarisation evaluation (non-crowdsourcing based) can be found in (Louis and Nenkova, 2013) and on Sentiment Analysis in (Balahur et al., 2014).

Argumentation Mining has gained increased interest in recent years, fuelled by annotated corpora becoming available (Palau and Moens, 2008; Walker et al., 2012; Stab and Cardie, 2014) and work spanning from classification of argumentative propositions in on-line user comments using SVMs (Park and Cardie, 2014), to analysing multilogue in order to classify relations between comments (Ghosh et al., 2014) and even using Textual Entailment in identifying agreement relations in discourse fora (Boltuzic and Šnajder, 2014).

5 Conclusion

In this paper we presented the Online Forum Summarisation (OnForumS) pilot task at MultiLing'15. OnForumS is a first attempt at encompassing automatic summarisation, argumentation mining and sentiment analysis into one shared task. It is also a pioneer in bringing crowdsourcing to the evaluation of systems for automatic summarisation and argument structure parsing.

We presented the evaluation strategy followed and the performance results for the participating systems.

We see two key challenges ahead: a more immediate one is to aggregate better the crowdsourcing data by using a probabilistic model of annotation (Passonneau and Carpenter, 2013), and a longer-term one is to bring in into the task definition higher-level units, such as whole interaction threads.

References

Alexandra Balahur, Eric van der Goot, Ralf Steinberger, and Andrés Montoyo, editors. 2014. *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*, Baltimore (MD), USA.

Braninimir Boguraev and Christopher Kennedy. 1997. Salience-based content characterisation of text documents. In Inderjeet Mani, editor, *Proceedings of the Workshop on Intelligent and Scalable Text*

Table 4: Results in terms of precision, recall and F1: English (top scores in bold).

	In_Favour			**Against**			**Positive**			**Negative**		
GroupAndRun	**R**	**P**	**F1**	**R**	**P**	**F1**	**R**	**P**	**F1**	**R**	**P**	**F1**
BASE-first	7.48	28.27	11.31	2.46	6.01	3.35	24.43	22.99	21.97	1.40	2.28	1.68
BASE-overlap	2.26	35.02	4.18	1.07	19.26	1.90	8.27	39.22	12.76	0.65	9.50	1.22
CIST-run1	67.86	24.49	34.94	0.18	1.03	0.28	45.14	24.35	28.58	2.01	2.27	1.97
CIST-run2	**70.79**	**25.18**	**35.99**	0.18	1.17	0.32	**45.61**	**24.64**	**28.72**	2.01	2.47	2.00
JRC-run1	6.78	34.60	10.78	1.15	8.89	2.00	10.01	29.14	12.77	1.37	6.81	2.24
JRC-run2	9.91	31.11	14.39	0.89	4.60	1.43	12.34	26.57	15.36	1.09	4.70	1.64
USFD_UNITN-run1	0.52	43.89	3.34	**5.44**	**5.15**	**4.39**	13.24	26.86	18.93	**3.00**	**5.83**	**6.21**
USFD_UNITN-run2	0.12	50.00	1.18	1.92	3.97	2.44	7.46	29.19	14.50	1.41	4.64	5.59
UWB-run1	12.91	39.16	17.70	0.06	16.67	0.42	6.69	37.75	11.25	0.00	0.00	0.00
UWB-run2	13.78	21.00	14.97	0.06	8.33	0.42	7.26	18.60	9.28	0.00	0.00	0.00

Summarization at the Annual Joint Meeting of the ACL/EACL, Madrid.

Filip Boltuzic and Jan Šnajder. 2014. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58, Baltimore (MD), USA.

C. Callison-Burch. 2009. Fast, cheap, and creative: Evaluating translation quality using amazons mechanical turk. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'09)*, volume 1, pages 286–295.

G. Erkan and D. Radev. 2004. LexRank: Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research (JAIR)*.

Debanjan Ghosh, Smaranda Muresan, Nina Wacholder, Mark Aakhus, and Matthew Mitsui. 2014. Analyzing argumentative discourse units in online interactions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 39–48, Baltimore (MD), USA.

George Giannakopoulos and Vangelis Karkaletsis. 2013. Summary evaluation: Together we stand npower-ed. In *Computational Linguistics and Intelligent Text Processing*, pages 436–450. Springer.

Annie Louis and Ani Nenkova. 2013. Automatically assessing machine summary content without a gold-standard. *Computational Linguistics*, 39(2):267–300.

H. Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165.

Raquel Mochales Palau and Marie-Francine Moens. 2008. Study on the structure of argumentation in case law. In *Proceedings of the Conference on Legal Knowledge and Information Systems*, pages 11–20.

Raquel Mochales Palau and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Joonsuk Park and Claire Cardie. 2014. Identifying appropriate support for propositions in on-line user comments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38, Baltimore (MD), USA.

Rebecca J. Passonneau and Bob Carpenter. 2013. The benefits of a model of annotation. In *Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse*, pages 187–195, Sofia, Bulgaria, August.

R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. 2008. Cheap and fastbut is it good?: Evaluating nonexpert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'08)*, pages 254–263.

Ian Soboroff. 2010. Test collection diagnosis and treatment. In *Proceedings of the Third International Workshop on Evaluating Information Access (EVIA)*, pages 34–41, Tokyo, Japan, June.

Christian Stab and Claire Cardie. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING*, pages 1501–1510.

Peter Turney and M. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21.

Marilyn Walker, Jean Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *Proceedings of LREC*.